Between-speaker rhythmic variability is not dependent on language rhythm, as evidence from Persian reveals

Asadi, Homa ; Nourbakhsh, Mandana ; He, Lei ; Pellegrino, Elisa ; Dellwo, Volker

Abstract: Acoustic measures of speech rhythm based on the durational characteristics of consonantal and vocalic intervals (henceforth C- or V-intervals) as well as syllabic intensity reveal between-speaker variability. The evidence obtained so far is based on speakers of stressed-timed languages, which are assumed to have complex consonant clusters and a higher degree of vowel reduction. Speakers of stressed-timed languages might operate their articulatory organs in different ways due to the syllable complexity and vowel reduction. Complex consonant clusters are released differently, and vowel reduction tends to be produced more or less strongly depending on speakers. When a language lacks such features, it is possible that rhythmic variation between its speakers decreases. In the present study, we aimed at exploring between- and within-speaker rhythmic variability in Persian, an Indo-European language categorised as syllable-timed. Acoustic correlates of speech rhythm (%V, ΔV[ln], ΔC[ln], n-PVI-V) and articulation rate were obtained from two Persian corpora with different sources of within-speaker variability. In the first corpus, the source of within-speaker variability mainly comes from non-contemporaneous recording sessions, and in the second corpus, from different speech rates. Results revealed that there were significant differences between speakers in all investigated speech rhythm measures in Persian and %V best discriminated between speakers. This reveals that the lack of typical stress-time features does not affect between-speaker variability in speech rhythm.

DOI: https://doi.org/10.1558/ijsll.37110

Between-speaker rhythmic variability is not dependent on language rhythm, as evidence from Persian reveals

Abstract

Acoustic measures of speech rhythm based on the durational characteristics of consonantal and vocalic intervals (henceforth C- or V-intervals) as well as syllabic intensity reveal between-speaker variability. The evidence obtained so far is based on speakers of stressed-timed languages which are assumed to have complex consonant clusters and a higher degree of vowel reduction. Speakers of stressed-timed languages might operate their articulatory organs in different ways due to the syllable complexity and vowel reduction. Complex consonant clusters are released differently, and vowel reduction tends to be produced more or less strongly depending on speakers. When a language lacks such features, it is possible that rhythmic variation between its speakers decreases. In the present study, we aimed at exploring between- and within-speaker rhythmic variability in Persian, an Indo-European language categorized as a syllable-timed. Acoustic correlates of speech rhythm (%V, ΔV[ln], ΔC[ln], n-PVI-V) and articulation rate were obtained from two Persian corpora with different sources of within-speaker variability. In the first corpus, the source of within-speaker variability mainly comes from non-contemporaneous recording sessions, and in the second corpus, from different speech rates. Results revealed that there were significant differences between speakers in all investigated speech rhythm measures in Persian and %V best discriminated between speakers. This reveals that the lack of typical stress-time features does not affect between-speaker variability in speech rhythm.

Keywords: speaker idiosyncrasies, speech rhythm, forensic phonetics

1. Introduction

Verbal communication is an inseparable part of human social interactions. Everyday experience of communicating with a group of people has proven that we can easily recognize each other's voice. This daily experience points to the fact that a speaker's voice carries a tremendous number of idiosyncratic features, which enables listeners to identify a familiar voice or discriminate different unfamiliar voices.

Comparing speakers based on acoustic parameters is a very important task in forensic phonetics research and practice. In a typical forensic voice comparison task (henceforth, FVC), the voice samples of a criminal (the unknown or disputed sample) and a suspect (the known sample) are compared, in order to assess the probability of observing the evidence under the assumption that the same speaker has produced both known and disputed samples versus the probability of observing the evidence under the assumption that two different speakers have produced the speech samples (Rose, 2002). To achieve this goal, a wide variety of phonetic or acoustic features are used. However, acoustic parameters differ in their suitability for forensic analyses. One of the chief criteria for suitability is that they have to exhibit high between-speaker variability and low within-speaker variability (Wolf, 1972; Nolan, 1983; Rose, 2002; Jessen, 2008; Morrison, 2010). Differences in speakers' voices stem from two sources, organic and learned differences (Wolf, 1972). Therefore, it is conceivable that acoustic parameters which originate from a close interaction between these two sources have great potential to show

variations across speakers. At the same time, acoustic parameters which are robust to different sources of within-speaker variability are of great importance in forensic phonetics. As Rose (2002: 24) states "it is a truism of phonetics that no-one ever says the same thing in exactly the same way"; the voices of the same speaker also vary when he/she is speaking on different occasions. Within-speaker variability comes from various sources such as non-contemporaneity of speaking occasions, emotional and health state, communication context and phonological environment (Nolan, 1983; Rose, 2002). Therefore, knowing features which fulfill the two criteria of high between-speaker variability and low within-speaker variability can contribute strongly to forensic speaker identification.

The forensic phonetics literature is replete with studies focusing on frequency domain features like fundamental frequency (e.g., Rose, 2003; Kinoshita, 2005; Lindh, 2006; Hudson et al., 2007), formant frequencies (e.g., Goldstein, 1972; Kinoshita, 2002; Nolan and Grigoras, 2005; Rose, 2007; Kahn et al., 2011; Gold et al., 2013) and spectral envelope features (e.g., Amino and Arai, 2009; Schindler and Draxler; 2013; Gordon et al., 2002). However, speakers' voices not only vary considerably in terms of their spectral characteristics but also by their rhythmic characteristics. In recent years, there has been an increasing interest to use temporal features of speech as an acoustic cue for forensic speaker identification. The rationale behind this idea is motivated by the fact that speakers have an individual anatomy of the articulatory apparatus which consequently leads to an individual way of steering them. It has been argued that this should have some impact on the properties of speech rhythm (Dellwo et al., 2007; Dellwo et al., 2015). Complexity of this relationship increases when language-specific rhythmic characteristics are taken into consideration. Numerous different studies on a large variety of languages have been carried out to assess the effectiveness of acoustic rhythm parameters in segregating between speakers. Yoon (2010) applied rhythm metrics on the ten speakers of the Buckeye corpus to investigate between- and within-speaker rhythmic variability. %V and VarcoV were found to be highly significant between speakers in this study. Wiget et al. (2010) investigated the robustness of various rhythm metrics to between-speaker variation based on 6 British English speakers. They showed a high degree of variability between speakers. Additionally, they found a large variability between utterances of varying lexical content. With the same objective, Leemann et al. (2014) also explored between-speaker variability of suprasegmental temporal features in a fully controlled corpus consisting of 16 speakers of Zurich German. Their study revealed a high degree of between-speaker variability in both read and spontaneous speech. They also found that suprasegmental temporal features are robust to speaking style and channel variability which consequently adds to their efficiency in application in forensic casework. Dellwo et al. (2015) found strong and consistent between-speaker variability in durational metrics of %V, ΔC(ln), ΔV(ln) and Δpeak(ln) in two corpora of German and Swiss German speakers. Their study showed that between-speaker rhythmic variability was robust against prosodic and linguistic sources of within speaker variability, such as rate differences, and individual lexical and morphosyntactic choices. They concluded that idiosyncratic articulatory movement is the crucial factor underlying existent rhythmic variation across speakers. More recently, researchers focused on intensity-based metrics of speech rhythm (He and Dellwo, 2014, 2016) claiming that individual movements of the articulators could possibly give rise to a great amount of speaker-specificity encoded in intensity aspects of speech signals. He (2018) further found that intensity-based rhythm measures successfully captured the articulatory differences between age groups, confirming that a strong association between articulation and intensity variability exists.

Evidence from previous studies supports the view that durational characteristics of C- and V- intervals vary strongly and significantly between speakers (Dellwo et al., 2015). Yet, much of the current literature on between-speaker rhythmic variability has largely focused on stressed-timed languages like English and German, whose speakers might vary considerably

in the way they operate the articulators to produce complex phonotactics and vowel reductions (Dellwo et al., 2015; Leemann et al., 2014; He and Dellwo, 2016). However, it is unknown whether speakers of languages classified as syllable-timed also vary in terms of their rhythmic characteristics the same extent speakers of stress-timed languages do.

Why do we expect a difference in between-speaker rhythmic variability in syllable-timed languages? Languages of different rhythmic classes are characterized by different phonological phenomena influencing consonantal and vocalic duration. Syllable-timed languages are known to have lower durational variability of both C- and V-interval duration as well as less complex syllable structure compared to stressed-timed languages (Roach, 1983; Ramus et al., 1999; Grabe and Low, 2002; Dellwo, 2010). Based on previous studies durational characteristics of C- and V- intervals are highly influenced by the phonology of a language. Prieto et al. (2012) showed that the complexity of syllables has also an impact on systematic variability in measurements of speech rhythm. Phonotactic structures thus influence measurable speech rhythm characteristics. Speakers behave differently in the particular form of gestures they use to implement a given underlying sound system (Dellwo et al., 2007). It thus seems plausible that a simpler phonotactic structure allows for less between-speaker variability. Complex consonant clusters can be released differently between speakers and vowel reduction can be carried out more or less strongly depending on the speaker. When such features are missing it is possible that the degree of between speaker durational variability decreases. We thus hypothesize that there would be little to no between-speaker variability in rhythm in syllable- as opposed to stress-timed languages. Here, we tested this hypothesis by studying rhythmic variability in Persian.

Persian belongs to the Indo-Iranian branch of the Indo-European language family and is the official language of Iran. Standard Contemporary Persian is the variety spoken by educated people in Tehran and in the media. Persian is categorized as a syllable-timed language (Windfuhr, 1979; Lazard., 1992) because it has a simple syllable structure of CV(C)(C) and according to most authors there is no vowel reduction pattern in Persian of the kind found in stress-timed languages. (Windfuhr, 1979; Lazard., 1992; Sheikh Sangtajan and Bijankhan., 2010; Yavaş, 2011; Sadeghi, 2015). Syllable-initial consonant clusters do not occur in Persian and syllable-final consonant clusters take maximally two consonants in their structure patterns. To date, very little attention has been paid to Persian in the realm of forensic phonetics. Moreover, among the very few studies, the primary focus was on segmental features. This study sets out to explore between- and within-speaker rhythmic variability in Standard Contemporary Persian. The primary goal of this research is to identify acoustically measurable speech rhythm parameters that have potential for forensic speaker identification in Persian. Another important purpose of this study is to contribute to forensic speaker comparison research in Persian with potential applications in forensic speaker comparison casework. The present study aims to answer the following questions:

1) Does acoustically measurable speech rhythm vary between speakers in Persian?

2) Which rhythmic metrics explain possible between-speaker variability best?

3) Which rhythmic metrics are most robust to sources of within-speaker variability?

These questions were studied by means of an explorative corpus analysis of two Persian speech corpora. Given that temporal acoustic parameters are sensitive to lexical choices (Wiget and et al., 2010, Dellwo et al., 2015), we used read speech in which people uttered lexically identical utterances. We also controlled extraneous factors e.g. age and accent to minimize their effect. In the first experiment, between-speaker variability of speech rhythm was studied in a database in which the source of within-speaker variability was time-lapsing between recordings

(henceforth: non-contemporaneous corpus). In the second experiment, we analyzed between-speaker rhythmic variability and strongly varying within-speaker speech rate (henceforth: Tempo corpus).

2. Method

2.1 The corpora

2.1.1 Non-contemporaneous corpus

24 speakers of Standard Contemporary Persian (12 male, 12 female) were recorded on 2 non-contemporaneous sessions, separated by a time-lapse of one to two weeks. The participants were aged between 22 and 35 years old (mean = 28.6, sd = 4.1). None of them reported any speech or hearing disorders. Speech materials consisted of read utterances based on 54 Persian sentences with average number of 11 syllables per sentence. Total duration of utterances was around 102 minutes. The dataset comprised 2592 tokens: 24 speakers × 54 sentences × 2 repetitions. Speakers were asked to read the 54 sentences one by one, with a pause and in a natural way, without any marked intonation. The microphone was positioned approximately 20 cm away from the mouth of the speakers in a diagonal position. Recording sessions were carried out in the sound proof booth at the phonetics laboratory of Alzahra University with the sampling rate of 44.1 kHz and a quantization of 16 bit.

2.1.2 Tempo corpus

10 native speakers of Standard Contemporary Persian (5 male, 5 female) were instructed to read *The North Wind and the Sun* in Persian at five different speech rates (normal, slow, slower, fast and fastest possible). The procedure was similar to the construction of the BonnTempo corpus (Dellwo, 2010). The participants were aged between 28 and 37 years old (mean = 32.2, sd = 3). The speech material contained 193 phonological syllables. The passage was subdivided roughly in 8 sentences. The dataset comprised 400 tokens (10 speakers × 8 sentences × 5 speech rates). Prior to each recording session, the participants were asked to read the text several times to familiarize with the passage. Subsequently, speakers were asked to slow down their rate (intended speech rate) in two steps and following that they were asked to read the text faster and as fast as they can. This created strong syllable rate variability in the five different reading passages. For data on German rate variability using this method see Dellwo et al. (2015). The recording venue and set-ups were the same as in Sec 2.1.1.

2.2 Segmentation and calculation of rhythm measures

Speech tokens were analyzed using Praat (version 5.2.34, Boersma and Weenink, 2013). The two speech corpora were annotated in three tiers: segments, CV-intervals and syllables. Firstly, segments on- and offsets were labeled manually using Praat's annotation function. Then CV intervals were created automatically using an automatic script *CV Creator Tier*[1]. A C-interval consists of one or more consonants preceded and followed by a vowel or by a pause whereas a V-interval consists of one or more vowels (or vocalic segments like diphthongs, triphthongs, etc.) preceded and followed by a consonant or by a pause (Dellwo, 2010). The syllable tier was also labeled manually by trained phoneticians. A Praat script *duration*Analyzer.praat[2] was used to automatically calculate the rhythm measures described in Sec 2.3.

2.3 Selection of acoustic parameters

Acoustic correlates of speech rhythm are based upon different phonetic durational units extending from CV-intervals (Ramus et al., 1999; Grabe and Low, 2002) over syllables or feet

(Nolan and Asu, 2009), voiced and unvoiced intervals (Dellwo and Fourcin, 2013) to amplitude peak intervals (Marcus, 1981). Such rhythmic measures also belong to two domains pertinent to durational characteristics of speech and amplitude envelope. For this study, we selected duration-based measures retrieved from CV-intervals and syllable units of speech signals. Since speech rate is potential to produce artefacts in obtaining the result, we only applied rate-normalized measures. In total we have chosen five duration-based measures as follows: one vocalic duration ratio (%V), two consonantal and vocalic duration variability measures (ΔV(ln), ΔC(ln)), one rate-normalized vocalic variability measures (n-PVI-V) and one rate measure based on syllable (articulation rate). Based on well-established temporal measures in the field of speech rhythm (Ramus et al.,1999; Grabe and Low, 2002; Dellwo et al., 2012; Leeman et al., 2014), acoustic rhythmic metrics were employed on the collected databases. From the CV interval tier, we automatically calculated the following duration-based measures: (Dellwo et al., 2015; Leemann et al., 2014):

-%V: proportion over which speech is vocalic;
- ΔV(ln): standard deviation of the natural-log normalized durations of vocalic intervals
-ΔC(ln): standard deviation of the natural-log normalized durations of consonantal intervals
-n-PVI-V: rate-normalized averaged durational differences between consecutive vocalic intervals;

From the syllable tier, we calculated the articulation rate (number of syllables per second). Speech rate is one of the best ranked discriminant parameters in forensic speaker comparison (Gold and French, 2011). Since articulation rate is based on linguistic units of syllables, it is assumed to be a good indicator of perceived speech rate (Dellwo, 2010), we thus calculated the articulation rate based on the number of syllables per second.

Above-mentioned acoustic rhythmic parameters are calculated via the following formulas:

$$\%\mathrm{V}=\frac{\left(\sum_{i=1}^{N_V} V_i\right)\cdot 100}{\sum_{i=1}^{N_C} C_i+\sum_{i=1}^{N_V} V_i}$$

where N_V is the total number of sampled V-intervals, N_C the total number of sampled C-intervals, V_i the duration of the *i*th V-interval and C_i is the duration of the *i*th C-interval.

$$\Delta\mathrm{C}\ln=\sqrt{\frac{N_\mathrm{C}\sum_{i=1}^{N_\mathrm{C}}(\mathrm{Ln}\ \mathrm{C}_i)^2-\left(\sum_{i=1}^{N_\mathrm{C}}(\ln \mathrm{C}_i)\right)^2}{N_\mathrm{C}(N_\mathrm{C}-1)}}$$

where N_C is the number of C-intervals in utterance and C_i is the duration of the *i*th C-interval.

$$\Delta\mathrm{V}\ln=\sqrt{\frac{N_\mathrm{V}\sum_{i=1}^{N_\mathrm{V}}(\mathrm{LnV}_i)^2-\left(\sum_{i=1}^{N_\mathrm{V}}(\ln \mathrm{V}_i)\right)^2}{N_\mathrm{V}(N_\mathrm{V}-1)}}$$

where N_V is the number of V-intervals in utterance and V_i is the duration of the *i*th V-interval.

$$\mathrm{nPVI}-\mathrm{V}=100\ \times\ \frac{\sum_{k=1}^{m-1}\left|\frac{d_k-d_{k+1}}{(d_k+d_{k+1})/2}\right|}{m-1}$$

where *m* is the number of vocalic intervals and d_k is the duration of the *k*th interval.

$$\text{articulation rate} = \frac{N_{syll}}{d}$$

where N_{syll} is the total number of syllables in an utterance and *d* is the total duration of the utterance in seconds (excluding pauses).

2.4 Statistical analyses

Statistical analysis of data was carried out using R (R core Team Year) version 3.3.3 and the R package *lme4* (Bates et al., 2016). Linear mixed-effects models were performed to analyze the significance of the between- and within-speaker variability on the selected acoustic rhythmic measures in the two speech corpora. To test the significance of an effect, two models are formed. In the first experiment, speaker and repetition were entered into the model as fixed effect and sentence was treated as random effect and in the second experiment speaker and speech rate were considered as fixed and sentence as random effect. In the full model, the effect in question was considered as either a fixed or a random effect [R code: full_model = (lmer (dependent_variable~fixed_factor+(1|random-factor)), data=data)] and in the reduced model, the effect in question was excluded [R code: reduced_model = (lmer (dependent_variable ~ 1 + (1|random-factor)), data=data)]. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the reduced model without the effect in question (R code: anova (full_model, reduced_model)).

Next, to address the second question of the research that which rhythmic metrics could account better for between-speaker variability, we employed a multinomial logistic regression model using SPSS (IBM Corp. 2012) to test which of the selected acoustic rhythmic parameters perform better in speaker identification. Speaker was considered as the nominal response variable and selected acoustic parameters were treated as the predicting variables. The amount of between-speaker variability explained by each durational measure was calculated via the likelihood ratio χ^2 of each acoustic parameter divided by the sum of likelihood ratio χ^2s of all parameters.

3. Experiment 1

3.1 Results

Table 1 contains the results of the linear mixed-effect model (fitted by maximum likelihood) for variation of the durational measures across speakers in the non-contemporaneous corpus of read speech. Figures 1-5 show the boxplots illustrating the between-speaker rhythmic variability across speakers on the two different occasions.

Table 1: Results of the linear mixed-effect model for duration-based measures

Measures	Factor tested	χ^2(df)	Result
%V	Speaker	1439.5 (23)	P < 0.0001
ΔC(ln)	Speaker	349.2 (23)	p < 0.0001
ΔV(ln)	Speaker	257.28 (23)	p < 0.0001
articulation rate	Speaker	2417.6 (23)	p < 0.0001
n-PVI-V	Speaker	141.66 (23)	P < 0.0001

We performed linear mixed effect-model for each rhythmic measure, comparing models with speaker and gender as fixed effect and sentence as random effect. The result showed no main effect of gender on all investigated rhythmic measures. Therefore, gender is excluded from the linear mixed-effect model for the following steps of statistical analyses.

As is obvious from the results shown in table 1, comparison between the full models and reduced models indicates a significant difference between the two models and all full models showed an increased goodness of fit which implies that between-speaker variability is significant for the measures of %V, ΔV(ln), ΔC(ln), n-PVI-V, and articulation rate in the non-contemporaneous corpus of read speech.

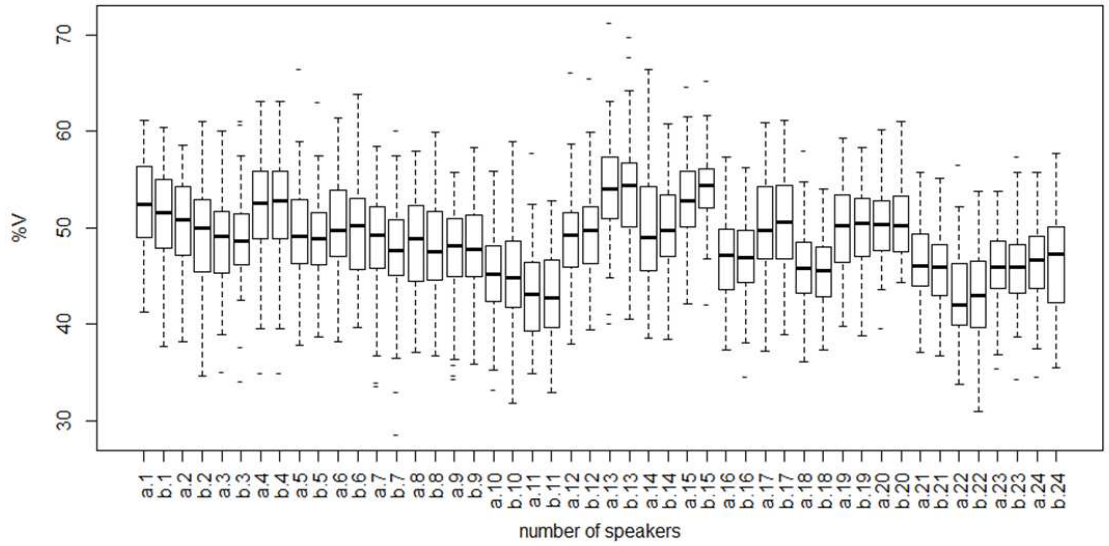

Figure 1: Boxplots of between- and within-speaker variability for the rhythmic measure of %V

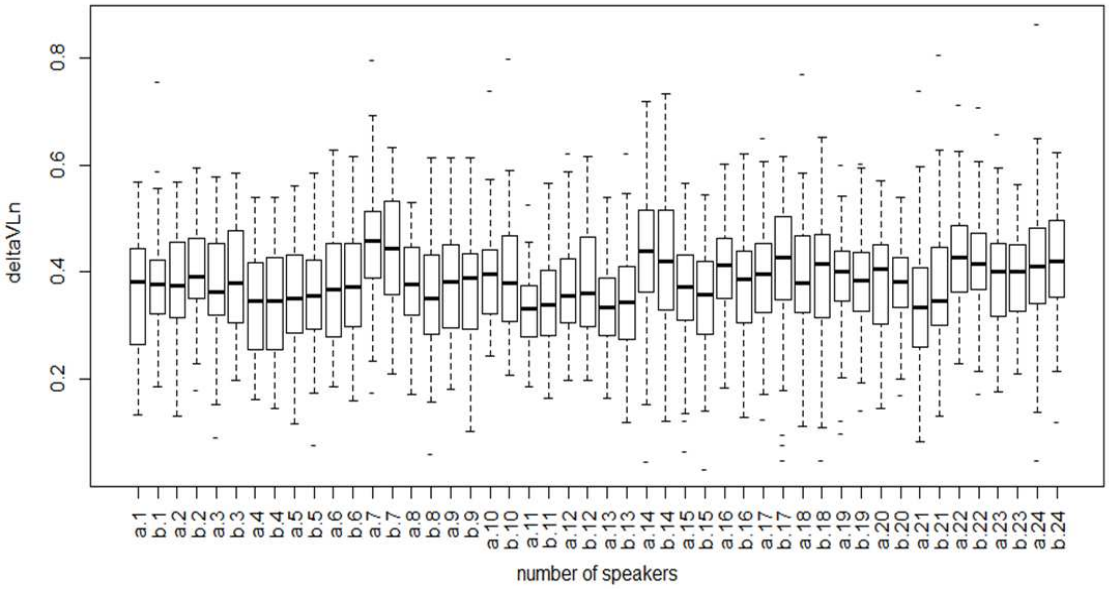

Figure 2: Boxplots of between- and within-speaker variability for the rhythmic measures of ΔV(ln)

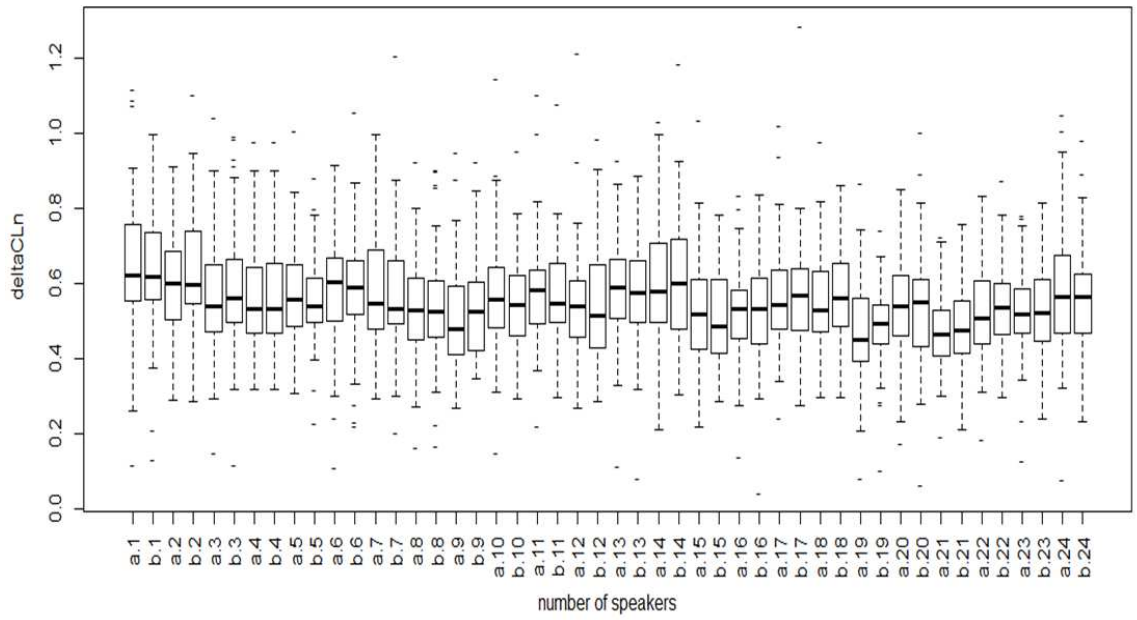

Figure 3: Boxplots of between- and within-speaker variability for the rhythmic measures of ΔC(ln)

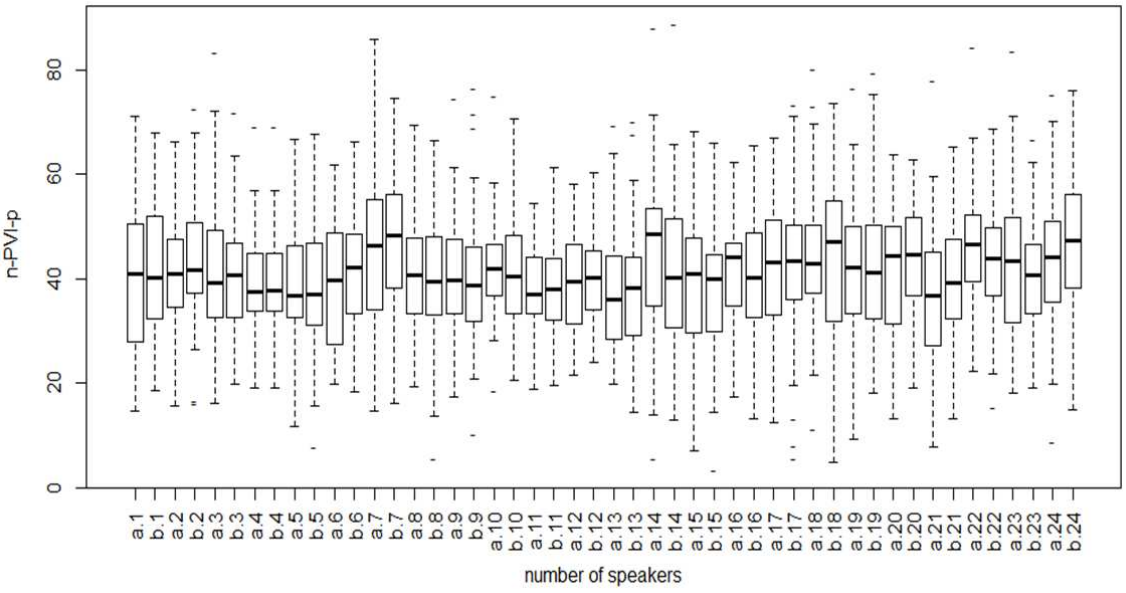

Figure 4: Boxplots of between- and within-speaker variability for the rhythmic measure of n-PVI-V

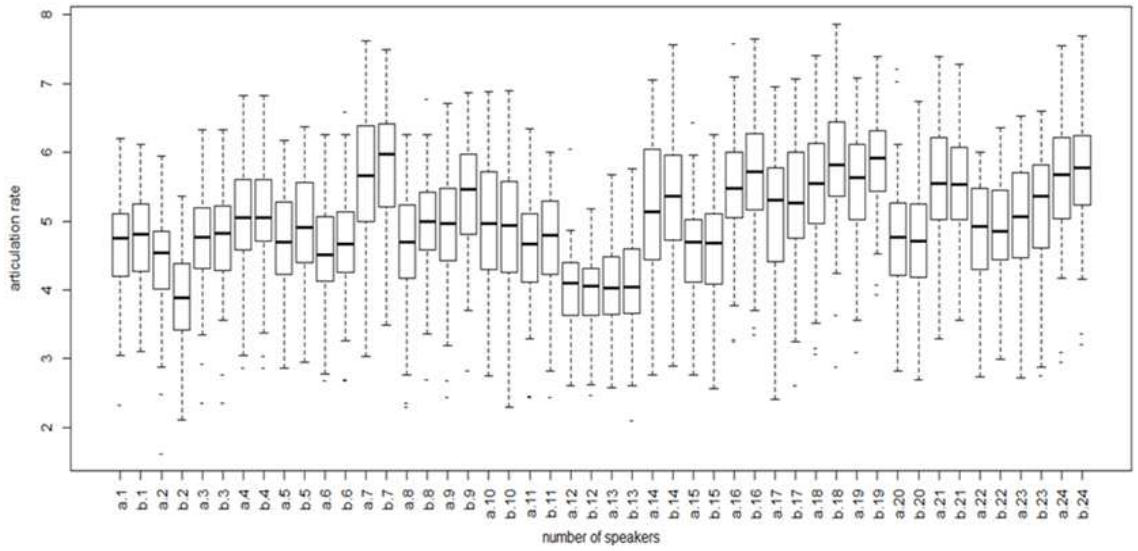

Figure 5: Boxplots of between- and within-speaker variability for the rhythmic measure of articulation rate

It is evident from figures 1-5 that some speakers could be distinguished well but for others the distributions overlap. For example, speakers 2 and 12 were similar in terms of %V while they vary regarding ΔC(ln). Conversely, speakers 3 and 18 may not be distinguished on ΔC(ln), but they can be differentiated on %V. This suggests that using a combination of variables may yield better results in speaker identification. Post hoc analyses were implemented using Bonferroni adjusted pairwise t-test to quantify the number of significant comparisons between speakers and to identify pairwise differences [R function: pairwise.t.test (data$dependent variable, data$speaker, p.adj="bonferroni")]. For %V, 334 out of the 529 (63%) possible paired comparisons were significant ($p < 0.0001$). Most speakers were differentiated by %V; however, speakers 2, 5, 6, 12, 14 and 19 were among those who showed similar scores. For ΔC(ln): 102 of the 529 comparisons (19%), ΔV(ln): 96 of the 529 comparisons (18%), n-PVI-V: 52 of the 529 comparisons (10%) and for articulation rate 328 of the 529 comparisons (62%) were significant ($p < 0.0001$). As for ΔV(ln) and n-PVI-V, speaker 7 has shown the greatest variation compared with other speakers. It means that speaker 7 behaved more differently in terms of vocalic durational variability while other speakers, in particular speakers 2 and 20, revealed very similar scores on these two measures. It can be inferred from the results that the highest number of significant between-speaker comparisons can be obtained with %V and articulation rate. Although there are significant differences between the speakers on all the metrics, post hoc paired comparisons revealed that only a few speakers could be differentiated by ΔC(ln) and ΔV(ln) and n-PVI-V.

To explore whether speakers have behaved differently or similarly on two different occasions, we tested the significance of repetition on each selected parameter. The results of within-speaker occasion-to-occasion variability analysis showed that the variability of all tested parameters as a function of repetition is not significant ($p>0.05$) except for articulation rate (χ^2 [1]= 15.367, $P <0.0001$). It subsequently indicates that in terms of %V, ΔC(ln), ΔV(ln) and n-PVI-V, speakers have behaved consistently on the two different occasions and aforementioned duration-based rhythmic measures are robust to the within-speaker occasion-to-occasion variability. But they showed different values for the articulation rate on the two

separate sessions which means that articulation rate is not robust against within-speaker variability with the source of time-lapsing.

To test which acoustic rhythmic measures accounted best for between-speaker variability, we applied a multinomial logistic regression model. Rhythm measures were predictor variables while speaker is treated as nominal response variable. To obtain the variability explained by each measure in percentage, we took the percentage of the χ^2 value of each measure over the sum of all χ^2 values for all measures. As the result in table 2 shows the strongest effects were found for %V (χ^2[23]= 1039.478, P <0.0001) and articulation rate (χ^2[23]= 1034.274, P <0.0001) explaining 43.38% and 43.16% of the variability between speakers respectively. Figure 6 displays the relative importance of each investigated acoustic parameter towards explaining between-speaker variability. The radius of each durational measure is proportional to its relative contribution in showing variability between speakers.

Table 2: Summary of the results of multinomial logistic regression for duration measures

Measures	-2 Log Likelihood of Reduced Model	χ^2(df)	P	Variability explained
%V	15335.616	1039.478 (23)	< 0.0001	43.38%
deltaCln	14487.039	190.901 (23)	< 0.0001	7.92%
deltaVln	14392.146	96.008 (23)	< 0.0001	4%
articulation rate	15330.412	1034. 274 (23)	< 0.0001	43.16%
n-PVI-V	14331.476	35.338 (23)	< 0.0001	1.47%

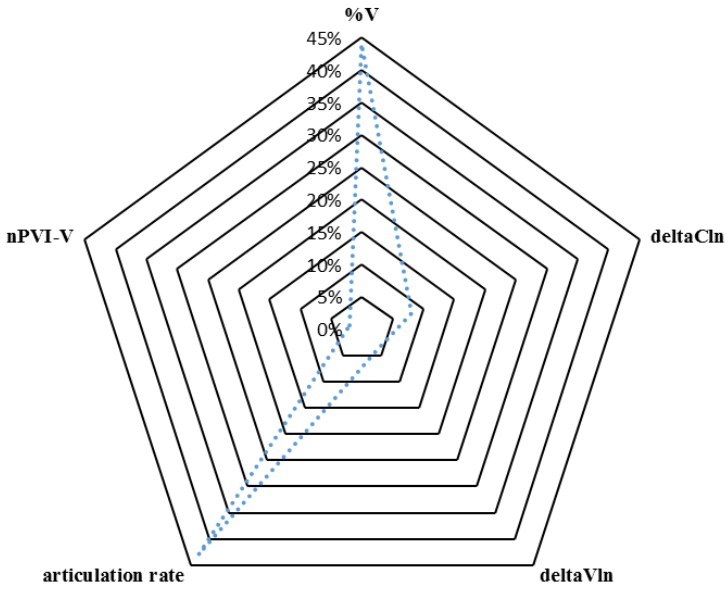

Figure 6: Radar chart illustrating the individual contribution of each investigated parameters in the multinomial logistic regression model for the nominal response variable 'speaker'.

3.2 Discussion

Results from experiment 1 replicated previous studies suggesting that durational characteristics of speech signals vary strongly and consistently between speakers (Dellwo et al., 2015; Leemann et al., 2014). As we mentioned before, most of research up to now has been restricted to analyzing stressed-timed languages like English and German and thus far no research has been carried out on between-speaker rhythmic variability of syllable-timed languages. Our hypothesis started from the premise that simple syllable structures allow speakers to have less leeway to vary in the production of the utterances. It is therefore expected that between-speaker variability might be reduced in syllable-timed-languages. However, our result revealed that there are strong and consistent differences between Persian speakers which in turn indicates that the lack of complex syllable structure as well as of vowel reduction do not affect the degree of rhythmic variability among speakers. This suggests that the between speaker differences in %V should be related to individual variability in the realization of vowel duration. Such individual variability might well arise from individual control mechanisms which determine the on- and offset of vowels.

%V and articulation rate showed the greatest number of differences between speakers respectively while n-PVI-V is the least useful parameter in segregating between speakers. This is in line with the findings of Wiget et al., (2010) who reported %V as a suitable parameter in segregating between English speakers. Moreover, in their study, no effect of speaker was obtained for n-PVI-V. In terms of %V, ΔC(ln) and ΔV(ln), similar results were also obtained in previous studies conducted on German (Dellwo et al., 2015) and Swiss German (Leemann et al., 2014) reporting strong between-speaker rhythmic variability and also emphasizing higher performance of %V at distinguishing between speakers. Despite the fact that ΔC(ln) and ΔV(ln) showed an effect of speaker, they are less discriminative than %V. Based on this result, we can conclude that regardless of the rhythmic typology of a language, %V encodes a considerable amount of speaker-specificity and thus has great potential- in distinguishing speakers in different languages (at least in languages investigated so far).

In terms of articulation rate, although strong variability was found between speakers, speakers were different in their two repetitions of the same sentences. This suggests that articulation rate can easily vary within the same speaker between sessions. This is plausible, as articulation rate is highly dependent on speakers' state, mood or emotion. It also suggests that the durational segment variability is less affected.

4. Experiment 2

4.1 Results

Since acoustically measurable speech rhythm is known to be highly affected by speech rate (Dellwo, 2010) we tested speech rhythm measures in a database in which speech rates varied strongly. Previous results on stress-timed German did not find for within-speaker variability as an effect of speech rhythm (Dellwo et al., 2015). Here we tested whether this finding holds for syllable-timed Persian. Table 3 provides the results obtained from the analysis of linear mixed-effect model (fitted by maximum likelihood) for variation of the durational measures between speakers across five different speech rates. Figures 7-11 also present the boxplots showing variability of rhythmic measures between speakers across different intended speech rates. Like experiment 1, no main effect of the factor gender was found on all investigated rhythmic measures. Therefore, gender is excluded from the statistical analyses.

Table 3: Results of the linear mixed-effect model for duration-based measures across different speech rates

Measures	Factor tested	χ^2(df)	Result
%V	Speaker	74.48 (9)	P < 0.0001
ΔC(ln)	Speaker	53.515 (9)	p < 0.0001
ΔV(ln)	Speaker	49.418 (9)	p < 0.0001
articulation rate	Speaker	154.13 (9)	p < 0.0001
n-PVI-V	Speaker	60.274 (9)	P < 0.0001

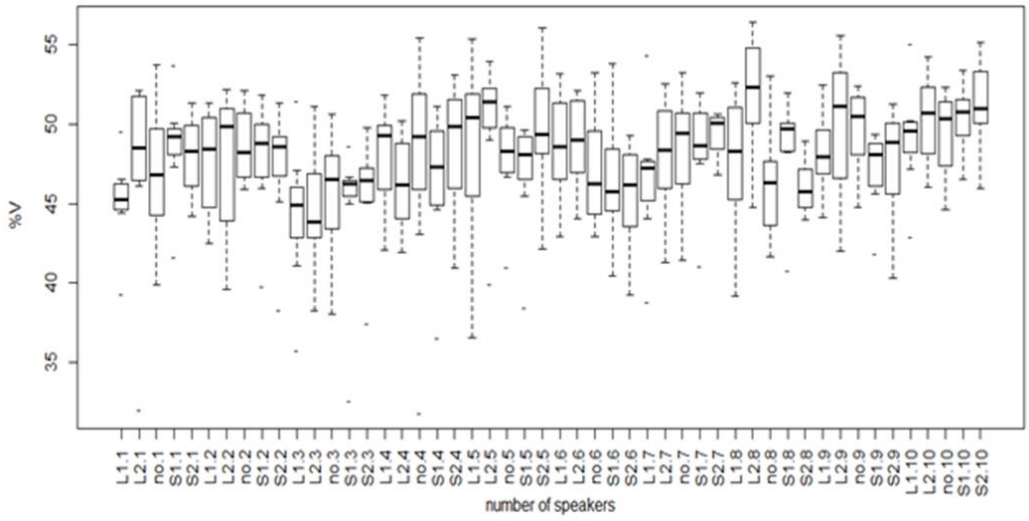

Figure 7: Boxplots of between- and within-speaker variability for the rhythmic measure of %V (speech rates are shown as follows: S2=fastest, S1=fast, no=normal, l2=slow, l1=slowest)

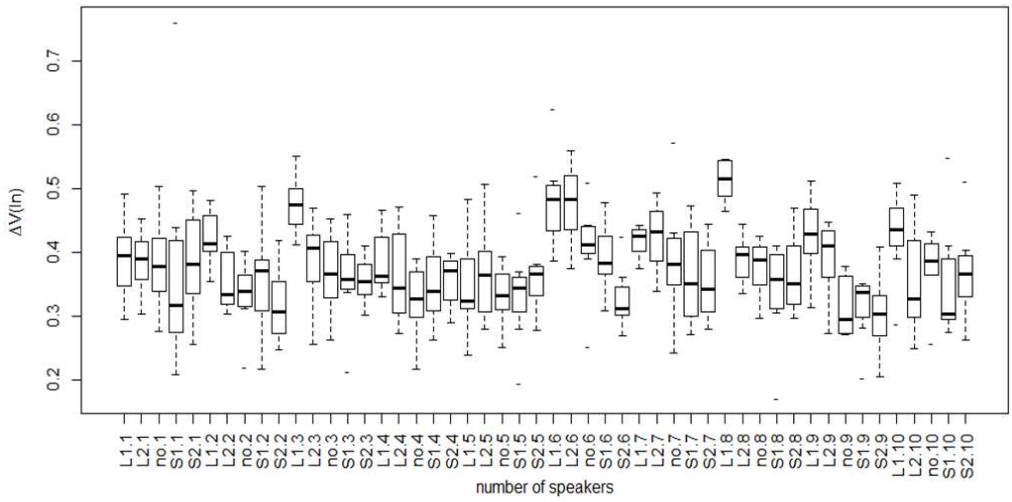

Figure 8: Boxplots of between- and within-speaker variability for the rhythmic measure of ΔV(ln) (speech rates are shown as follows: S2=fastest, S1=fast, no=normal, l2=slow, l1=slowest)

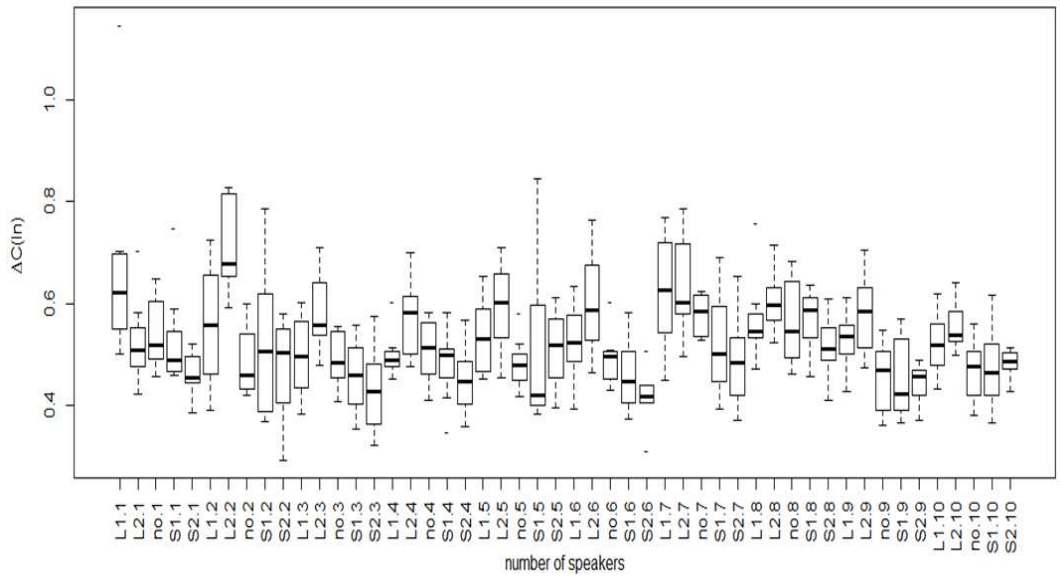

Figure 9: Boxplots of between- and within-speaker variability for the rhythmic measure of ΔC(ln) (speech rates are shown as follows: S2=fastest, S1=fast, no=normal, l2=slow, l1=slowest)

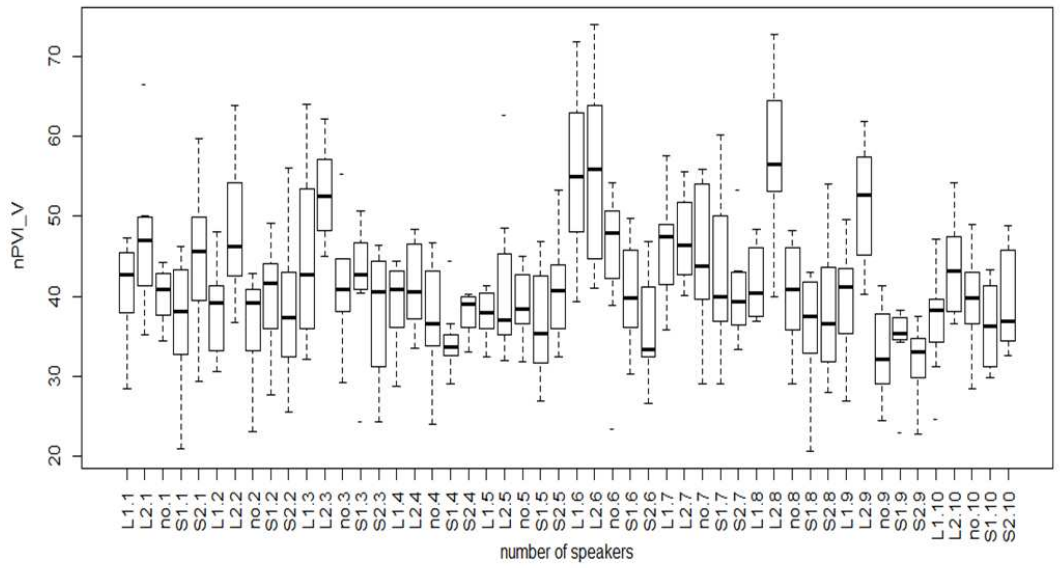

Figure 10: Boxplots of between- and within-speaker variability for the rhythmic measure of n-PVI-V (speech rates are shown as follows: S2=fastest, S1=fast, no=normal, l2=slow, l1=slowest)

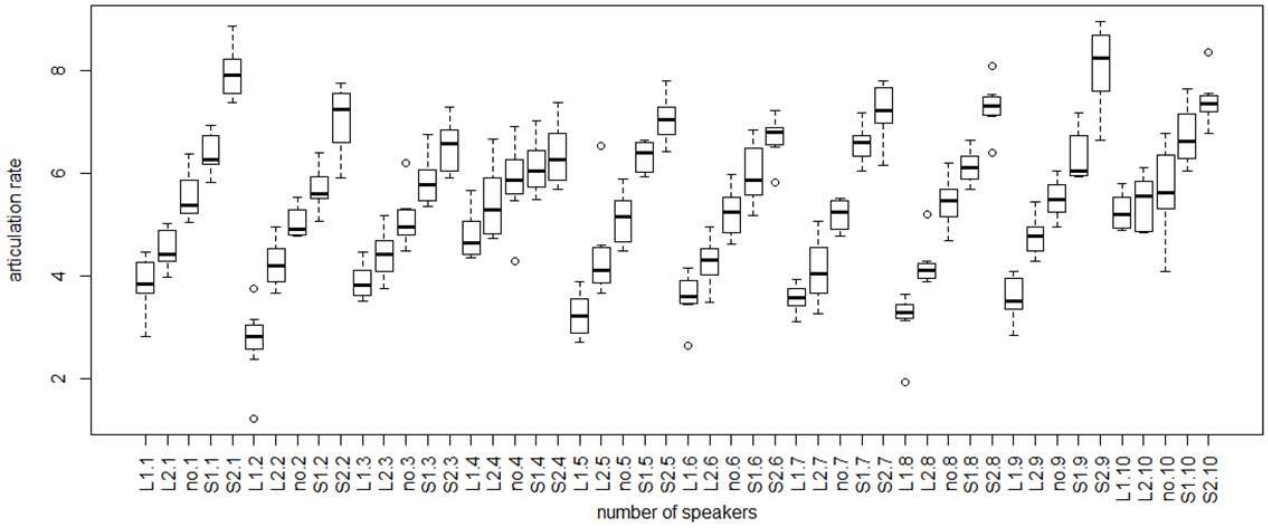

Figure 11: Boxplots of between- and within-speaker variability for the rhythmic measure of articulation rate (speech rates are shown as follows: S2=fastest, S1=fast, no=normal, l2=slow, l1=slowest)

Results revealed that the effect of speaker on all the rhythmic measures was significant across different speech rates. Post hoc analyses using Bonferroni adjusted pairwise t-test were applied separately for each speech rate to quantify the number of differences between speakers. For %V, 27 out of the 81 (33%) possible paired comparisons are significant ($p < 0.05$). For ΔC(ln): 10 of the 81 comparisons (12%), ΔV(ln): 8 of the 81 comparisons (10%), n-PVI-V: 9 of the 81 comparisons (11%) and for articulation rate 6 of the 81 comparisons (7%) are significant ($p < 0.05$). It is apparent from the results that %V appears to be contributing most to the discrimination of speakers, and that the between-speaker variability shown by the remaining measures is much smaller. If we compare with the results obtained in the first experiment, we can conclude that between-speaker variability is reduced in the dataset with high prosodic variability. It means that speech rate reduced the discriminatory power of the speech rhythm measures.

Considering the effect of speaker at each separate speech rate, we found that speakers are differentiated better when they speak at normal, slow and slowest rate. The results showed a main effect of speaker for %V and articulation rate at each of the five different intended speech rates ($P<0.0001$). As with ΔC(ln), the effect of speaker was not significant at fast rate (χ^2 [9]= 112.09, p=0.1362) and for ΔV(ln), it was not significant at fast rate ($\chi^2[9] = 6.927$, p=0.6446) and fastest possible rate ($\chi^2[9] = 17.413$, p=0.4263). For n-PVI-V, the effect of speaker was not significant at fastest rate ($\chi^2[9] = 17.941$, p=0.3579). We must therefore conclude that the discriminatory power of rhythmic measures declines with increasing speech rate.

To analyze within-speaker variability, the significance of articulation rate on each selected parameter was tested. The results showed that the within-speaker variability as a function of articulation rate is significant for ΔC(ln) ($\chi^2[1] = 112.09$, $P <0.0001$), ΔV(ln) ($\chi^2[1] = 44.769$, $P <0.0001$), n-PVI-V ($\chi^2[1] = 83.347$, $P <0.0001$), and articulation rate ($\chi^2[1] = 475.7$, $P <0.0001$), but not for %V ($\chi^2[1]= 1.0937$, P=0.2957). It means that speech rate variability within speakers did not have a significant influence on %V variability while it has strong effects on the remaining rhythm measures under investigation.

4.2 Discussion

Results from experiment 2 revealed that selected acoustics duration-based measures showed an effect of speaker in a dataset where prosodic within-speaker variability was very high. Despite the fact that the effect of speaker was significant for all the speech rhythm measures, their speaker-discriminatory power dramatically reduced. Except for %V, speakers showed a great dissimilarity toward themselves for the remaining parameters. Therefore, %V appears to be the most robust hence useful rhythmic parameter to capture between-speaker variability. We found that speech rate variability within speakers did not have a systematic influence on %V variability. In other words, %V remained stable between speakers even when within-speaker variability as a function of articulation rate was high. It is in line with the findings of (Dellwo et al., 2015) who also reported %V as a powerful speaker-specific parameter with robustness against different sources of variability.

Analysis of rhythmic parameters at each speech rate showed that the effect of speaker was significant at all normal rates while it was not significant for some parameters like ΔV(ln), ΔC(ln) and n-PVI-V at fast and fastest rate. This is contrary to the findings of Dellwo et al. (2015) who found significant between-speaker variability for the measures of ΔV(ln) and ΔC(ln) in the BonnTempo corpus containing speech with comparable rate variability. A possible explanation of the difference between Persian and German speakers might be related to the speed strategy that they have acquired. It seems possible that Persian speakers have used the same strategy in reading the passage at fast rate and that is why they are more similar in durational variability of consonantal and vocalic intervals at fast rate. In other words, the strategies Persian speakers use to accelerate their speech is similar, but when speaking at normal or slow rate the strategy taken by them involves changes in durational variability of consonantal and vocalic intervals. On the other hand, speakers of stress-time languages have less room for variation and thus behave more similarly across different speech rates.

In general, based on the results obtained from exploring different languages we can conclude that speech rate variability makes duration-based measures less discriminative. As it was mentioned in the introduction part, a parameter is considered forensically useful provided that it shows high between-speaker variability and high consistency within speakers (i.e. low within-speaker variability). In view of the results from both experiments carried out in this study, %V best fulfills both criteria. An advantage of %V over other parameters is its robustness against both sources of within-speaker variability tested in this study i.e. time-lapsing and speech rate variability. The discriminatory role of %V in forensic speaker comparison in stressed-timed languages also has been noted in prior studies.

5. Conclusion

In this study a selection of widely used rhythmic measures based on the durational characteristics of CV-intervals and syllables was applied to two Persian databases designed for exploring within- and between-speaker variability in speech. The focus of this study was mainly on discovering speaker idiosyncrasy in durational properties of speech signals produced by Persian speakers. Our findings could be summarized as follows:

- There is significant between-speaker rhythmic variability in Persian speakers.

- %V (the percentage over which speech is vocalic) yields the best result for speaker discrimination in both Persian datasets.

- %V is robust against within-speaker variability as a function of time.

- %V is robust against within-speaker variability as a function of articulation rate.

- Simpler syllable structure in Persian as well as lower degree of vowel reduction compared to stressed-timed languages like English and German do not have an influence on the degree of between-speaker rhythmic variability.

This study adds more evidence of speaker-individuality of suprasegmental temporal features. Taken together with previous research showing no impact of prosodic and linguistic factors on between speaker rhythmic variability, the outcomes of the present study provide further evidence for an articulatory explanation of between-speaker rhythmic variability. Our study also showed that language rhythm doesn't have an impact on between-speaker variability of speech rhythm measures. It seems that both stress-timed and syllable-timed languages have their own kind of linguistic demands and have about the same room for idiosyncrasy. A major finding of this study is that despite the typological difference across languages, %V comes out as the universal speaker characteristic that cuts across the typological difference.

Our findings pertaining to the potentials of using durational measures as a forensic cue may be of particular interest in the procedures where speaker identification information is needed. In future studies, additional potential speaker-specific rhythmic measures as well as intensity measures will be examined in order to determine a set of well-established discriminant rhythmic parameters in Persian. We will try to perform our analyses in a spontaneously produced speech dataset and telephone transmitted speech which is of relevance in forensic studies so as to assess the influence of speaking style and channel variability on between-speaker rhythmic variability in Persian.

References

Amino, K and Arai, T. (2009). "Speaker-dependent characteristics of the nasals". *Forensic Science International,* 185: 21-28.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2016). *lme4: Linear mixed-effects models using Eigen and S4* (R package version 1.1-7). http://CRAN.R project.org/ package=lme4, Accessed 24 November 2016.

Boersma, P. and Weenink, D. (2013) Praat: Doing Phonetics by Computer. http://www.praat.org, Accessed 13 July 2013.

Dellwo, V. (2010). *Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence*. PhD dissertation, Bonn University.

Dellwo, V. and Fourcin, A. (2013) "Rhythmic characteristics of voice between and within languages". *Travaux Neuchâtelois de Linguistique*, 59: 87–107.

Dellwo, V., Huckvale, M. and Ashby, M. (2007). "How is individuality expressed in voice? An introduction to speech production and description for speaker classification". In C. Müller (Ed), *Speaker Identification 1*, 1-20, Berlin: Springer Verlag.

Dellwo, V., Leeman, A. and Kolly, M. (2015). "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors". *The Journal of the Acoustical Society of America*, 137:1513-1528.

Dellwo, V., Leemann, A. and Kolly, M. (2012). "Speaker idiosyncratic rhythm features in the speech signal". In *Interspeech*, Portland, USA.

Gold, E., French, J.P and Harrison, P (2013). "Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework". In *Proceedings of Meetings on Acoustics*, Montreal, Canada, (pp. 1-8).

Gold, E., French, J.P (2011). "International practices in forensic speaker comparison". *The International Journal of Speech, Language and the Law*, 18(2), 293-307.

Goldstein, U. (1976). "Speaker-identifying features based on formant tracks". *The Journal of the Acoustical Society of America*, 59(3), 176-182.

Gordon, M, Barthmaier, P, Sands. K. (2002). "A Cross-linguistic study of voiceless fricatives". *Journal of the International Phonetic Association*, 32(2), 2-32.

Grabe, E. and Low, E. L. (2002). "Durational variability in speech and rhythm class hypothesis". In N. Warner & C. Gussenhoven (Eds.), *Papers in Laboratory Phonology* 7, 515-543, Berlin and New York: Mouton de Gruyter.

He, L. (2018). "Development of speech rhythm in first language: The role of syllable intensity variability". The *Journal of the Acoustical Society of America*, 143(6), 463-467.

He, L. and Dellwo, V. (2016). "The role of syllable intensity in between-speaker rhythmic variability". *The International Journal of Speech, Language and the Law*. Vol 23, 243-273.

He, L., and Dellwo, V. (2014). "Speaker idiosyncratic variability of intensity across syllables". In *Proceedings of INTERSPEECH*, (pp. 233-237), Singapore.

Hudson, T., de Jong, G., McDougall, K., Harrison, P., and Nolan, F. (2007). "F0 statistics for 100 young male speakers of Standard Southern British English". In *16th Proceedings of the International Congress of Phonetic Sciences*, Saarbrücken, (pp. 1809-1812).

IBM Corp. (2012). IBM SPSS Statistics for Windows (version 21.0). Armonk, NY: International Business Machines Corporation.

Jessen, M. (2008). "Forensic phonetics". *Language and Linguistics Compass,* 2(4): 671–711.

Kahn, J., Audibert, J.F.B., and Rossato, S. (2011). Inter and intra-speaker variability in French: An analysis of oral vowels and its implication for automatic speaker verification. *International Congress of phonetic sciences (ICPhS)*, 17(pp. 1002-1005).

Kinoshita Y. (2005). “Does Lindley’s LR estimation formula work for speech data? Investigation using long-term F0”. *The International Journal of Speech, Language and the Law*, 12(2): 235-254.

Kinoshita, Y. (2002). “Use of likelihood ratio and Bayesian approach in forensic speaker identification”. In *Proceedings of the 9th Australian International Conference on Speech Science and Technology.* Melbourne, Australia, (pp. 297-302).

Lazard, G. (1992). *Grammar of contemporary Persian*. Mazda Publishers.

Leemann, A., Kolly, M.-J., and Dellwo, V. (2014). “Speaker-individuality in suprasegmental temporal features: implications for forensic voice comparison”. *Forensic Science International*, 238, 59-67.

Lindh J. (2006). “Preliminary descriptive F0-statistics for young male speakers”. *Lund University Working Papers*, 52, 89-92.

Marcus, S. (1981). “Acoustic determinants of perceptual center (p-center) location”. *Perception and Psychophysics*, 30: 247–256.

Morrison, G. S. (2010). “Forensic voice comparison”. In I. Freckelton & H. Selby, *Expert Evidence*. Sydney: Thomson Reuters.

Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

Nolan, F. and Asu, E. L. (2009). “The pairwise variability index and coexisting rhythms in language”. *Phonetica*, 66(1–2): 64–77.

Nolan,F and Grigoras, C. (2005). “A case for formant analysis in forensic speaker identification”. *The International Journal of Speech Language and the Law*, 12(2): 143-173.

Prieto, P., del Mar Vanrell, M., Astruc, L., Payne, E., and Post, B. (2012). “Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish,” *Speech Communication*, 54, 681–702.

R Core Team (2014) R: A Language and Environment for Statistical Computing (version3.3.3). R Foundation for Statistical Computing. http://www.Rproject.org, Accessed 20 November 2016.

Ramus, F., Nespor, M. and Mehler, J. (1999). “Correlates of linguistic rhythm in the speech signal”. *Cognition*, 73, 265-292.

Roach, P. (1983). *English phonetics and phonology*, Cambridge: Cambridge University Press.

Rose, P. (2002). *Forensic speaker identification*, New York: Taylor & Francis.

Rose, P. (2003). “The technical comparison of forensic voice samples”. In I.S. Freckleton and H. Selby (eds.) *Expert evidence*. North Ryde: Lawbook Co, Ch. 99.

Rose, P. (2007). “Forensic speaker discrimination with Australian English vowel acoustics”. In *Proceedings of the 16th International Congress of Phonetic Sciences*. Saarbrücken, Germany, (pp. 1817-1820).

Sadeghi, V. (2015). “A phonetic study of vowel reduction in Persian”, *Language Related Research,* 30:165-187.

Sheikh Sangtajan, Sh. and Bijankhan, M. (2010). “The study of vowel reduction in Persian spontaneous speech”, *Journal of Research in Linguistics*, 2: 35-48.

Schindler, C., Draxler, Ch (2013). “Using spectral moments as a speaker specific feature in nasals and fricatives”. In *Proceedings of INTERSPEECH*, (pp. 2793-2796), Lyon, France.

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S. L. (2010). “How stable are acoustic metrics of contrastive speech rhythm?”, *The Journal of the Acoustical Society of America*, 127(3), 1559–1569.

Windfuhr, G. L. (1979). *Persian grammar: History and state of its study*. New York: Mouton de Gruyter.

Wolf, J. (1972). "Efficient acoustic parameters", *The Journal of the Acoustical Society of America,* Vol 51, pp 255-272.

Yavaş, M. (2011). *Applied English phonology*, United Kingdom: Wiley-Blackwell.

Yoon, T.J. (2010). “Capturing inter-speaker invariance using statistical measures of speech rhythm”. *In Electronic Proceedings of Speech Prosody*, (pp. 1-4), Chicago/IL, USA.

Appendix: Reading materials in the non-contemporaneous corpus

Some examples of the non-contemporaneous corpus are listed below in Persian, English translation and IPA:

۱) هر کسی حق داره راجع به این مسأله اظهارنظر کنه.

[har casi haɢ dɑre rɑdʒe be in masʔale ezhɑre nazar kone].

Everybody has the right to express his/her opinion about this issue.

۲) ماشین علی بژ هستش.

[mɑʃine ʔali beʒ hasteʃ].

Ali has a beige car.

۳) شایان چهارتا کلوچه خورد.

[ʃɑjɑn tʃɑhɑr tɑ kulutʃe xorɗ].

Shayan ate four cookies.

۴) مژده فرشته رو به خونه رسوند.

[moʒde fereʃtaro be xune resund].

Mozhde drove Fereshte to home.

۵) بچه‌های کشاورزی دنبال خرید بذر بودند.

[batʃehɑje ceʃɑvarzi dombɑle xaride bazr budand].

Agriculture students were looking to buy seeds.

۶) هر تیمی شهرآوردو ببره، می‌ره صدر جدول.

[har timi ʃahrɑvardo bebare mire sadre dʒadval].

Whichever team wins the derby game, it will occupy the first rank of the table.

۷) سگ و گربه به جون هم افتادند.

[sago gorbe be dʒune ham oftɑdan].

Dog and cat started fighting together.

۸) فضای خونه بدبو شده؛ هود رو روشن کنید.

[fazɑje xune badbu ʃode hud ro rowʃan konid].

The house is full of smells; turn on the ventilation hood.

۹) فالگیر زیر لب ورد خوند.

[fɑlʁir zire lab verd xund].

The fortuneteller muttered incantations.

۱۰) قاتل دیروز صبح یه نفرو کشت و فرار کرد.

[ɢɑtel diruz sob je nafaro koʃto farɑr kard].

The murderer killed someone yesterday morning and ran away.

[1] script is available under http://www.cl.uzh.ch.
[2] script is available under http://www.cl.uzh.ch.